Disease Symptoms and Patient profile Dataset Analysis using Predictive Models

Arya Sasi(L00179434)
MSc.Big Data Analytics
Atlantic Technological University
L00179434@atu.ie

***Abstract* — To advance medical research and enhance patient care, it is imperative in the field of healthcare to comprehend the complex interrelationships among symptoms, demographics, and health indicators.The dataset for this project is obtained from kaggle and consists of information related comprehensive compilation of symptoms and patient profiles for a wide range of diseases. The dataset is organized as a CSV file and includes a thorough collection of attributes.The main goals of this project are to answer the Questions(1) For which disease the outcome Variable is more positive(2)For which gender have high cholesterol level and Blood Pressure level and (3)Which Disease have the most positive and negative Outcome Variable.The data analysis process include carefully choosing and preparing data, as well as dealing with missing numbers.The goal is to improve decision-making by finding predictive indicators using machine learning models.This dataset has the potential to revolutionize our comprehension of healthcare. From the dataset we can compare different symptoms for same disease for different peoples also. This information may be utilized by a variety of stakeholders, including healthcare professionals, medical researchers, and healthcare technology companies.The following machine learning regression approaches were applied: (a) LightGBM Classifier (b) Random forest Classifier.**

Keyword-Diseases,Symptoms,Healthcare,Positive,Negative,LightGBMClassifier ,Random forest Classifier.

I. INTRODUCTION

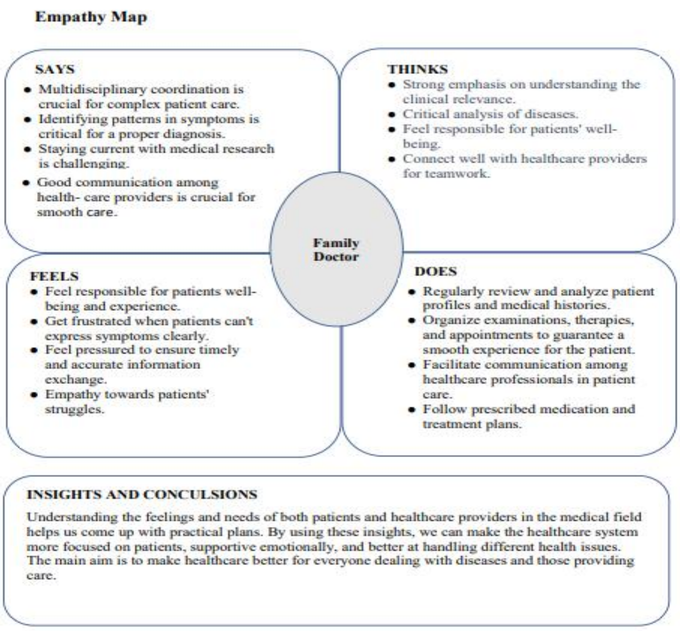
In the vast field of healthcare, understanding how symptoms, demographics, and health indicators connect is a puzzle we aim to solve using the Comprehensive Disease Symptom and Patient Profile Dataset. We want to explore this dataset to learn more about how symptoms like fever, cough, and fatigue relate to a person's age, gender, and health stats.

The goal is to find hidden patterns and unique profiles for different diseases. Whether you are a medical researcher, healthcare professional, or just curious about data, this collection offers valuable insights into health patterns. Uncover hidden trends, discover unique symptom profiles, and gain a deeper understanding of medical conditions. This dataset has the potential to revolutionize our comprehension of healthcare.

From the dataset we can compare different symptoms for same disease for different peoples also.This information can be used for clinical analysis, research studies, and epidemiological investigations pertaining to various diseases by physicians, researchers, and medical practitioners. Understanding the frequency and symptom patterns in people with particular medical disorders can be aided by it. This dataset can be used by various stakeholders, including:

- Healthcare Professionals
- Medical Researchers
- Healthcare Technology Companies.

An empathy map is a visual representation that helps to understand and empathize with their target users or customers. It's commonly used in design thinking processes to gain insights into users' needs, thoughts, emotions, and behaviors.So I contacted an Expert in Medical field to know more.The goal of this conversation was to gain insights into the experiences, perspectives, and needs of patients dealing with various medical conditions. The focus was on understanding not only the diseases but also the approaches and challenges faced by patients in managing their health.



Empathy Map

SAYS
- Multidisciplinary coordination is crucial for complex patient care.
- Identifying patterns in symptoms is critical for a proper diagnosis.
- Staying current with medical research is challenging.
- Good communication among health- care providers is crucial for smooth care.

THINKS
- Strong emphasis on understanding the clinical relevance.
- Critical analysis of diseases.
- Feel responsible for patients' well-being.
- Connect well with healthcare providers for teamwork.

Family Doctor

FEELS
- Feel responsible for patients well-being and experience.
- Get frustrated when patients can't express symptoms clearly.
- Feel pressured to ensure timely and accurate information exchange.
- Empathy towards patients' struggles.

DOES
- Regularly review and analyze patient profiles and medical histories.
- Organize examinations, therapies, and appointments to guarantee a smooth experience for the patient.
- Facilitate communication among healthcare professionals in patient care.
- Follow prescribed medication and treatment plans.

INSIGHTS AND CONCULSIONS

Understanding the feelings and needs of both patients and healthcare providers in the medical field helps us come up with practical plans. By using these insights, we can make the healthcare system more focused on patients, supportive emotionally, and better at handling different health issues. The main aim is to make healthcare better for everyone dealing with diseases and those providing care.

Figure 1 :Empathy Map

II.DATASET OVERVIEW

The dataset for this project is obtained from kaggle and consists of information related comprehensive compilation of symptoms and patient profiles for a wide range of diseases.This interesting dataset has a wealth of information, demonstrating fascinating links between symptoms and health indices. Explore the rich tapestry of fever, cough, weariness, and difficulty breathing, which are connected with age, gender, blood pressure, and cholesterol. The dataset enables users to reimagine our understanding of healthcare by giving profound insights and ground-breaking information. It's a breakthrough tool with the potential to transform how we approach and understand various medical conditions, making it a vital resource for anyone working to advance the area of healthcare. The dataset is organized as a CSV file and includes a thorough collection of attributes. The dataset mainly contain columns such as follows :

- Disease: The name of the disease or medical condition.
- Fever: Indicates the patient has a fever or not (Yes/No).
- Cough: Indicates the patient has a cough or not (Yes/No).
- Fatigue: Indicates the patient experiences fatigue or not (Yes/No).
- Difficulty Breathing : Indicatesr the patient has difficulty breathing (Yes/No).
- Age: The age of the patient.
- Gender: The gender of the patient (Male/Female).
- Blood Pressure: The blood pressure level of the patient (Low/Normal /High).
- Cholesterol Level: The cholesterol level of the patient (Low/Normal /High).
- Outcome Variable:Indicating the result of the diagnosis for the specific disease (Positive /Negative).

III. METHODOLOGY

The goal of this study is to find a Machine Learning model that can accurately analysis the Disease Symptoms and Patient Profile Dataset.A collection of comments and their related metadata, such as Disease and their Symptoms are input into the proposed system. After that, the data is turned into a features dataset, which is then used in the learning phase. Preprocessing is a transformation that entails a series of steps including cleaning, filtering, and encoding.

There are two sections to the preprocessed dataset: training and testing.Using the training dataset ,the training module builds a decision model that can be used to the test dataset. The model must be able to analyse the dataset. There are two sections to the dataset: training and testing. The model should be trained with more data to increase learning accuracy.The steps that require to be followed are:

- ➢ Data Collection
- ➢ Data Pre-processing
- ➢ Model Building
- ➢ Analyzing
- ➢ Result

A.Data Collection: The user has a dataset for analyzing, and the dataset's attributes are used to train the model. The dataset contains 349 rows and 10 columns. Headings and labels are the columns. The heading denotes Disease and their Symptoms of some patients that may or may not indicating the result of the diagnosis for the specific disease. This is to distinguish the result of the diagnosis or assessment for the specific disease is is more Positive / Negative

B.Data Pre-processing: According to the Disease Symptoms and Patient Profile Dataset, The attributes are divided into three groups: Boolean data, category data, and numerical data. Preparing the data is an important part of the pipeline for machine learning and data analysis. A format that is appropriate for analysis or model training is created by cleaning and converting raw data. Here are some key steps involved in data preprocessing:

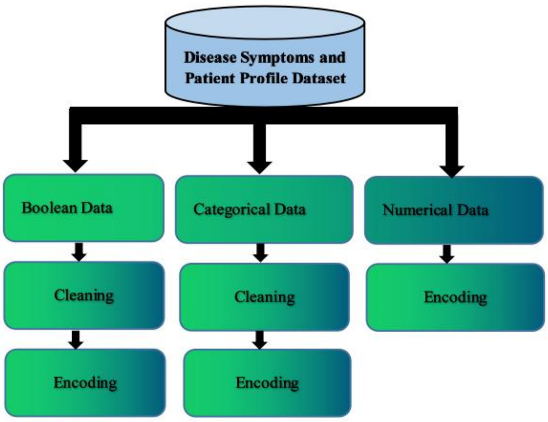


Figure 2 : Data Pre-processing.

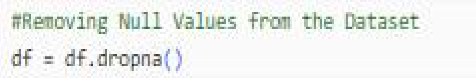
Data Cleaning:

- Handling missing values: Decide whether to remove or impute missing data based on the context.
- Removing duplicates:Eliminate duplicate entries to ensure data integrity.
- Handling outliers:Address extreme values that may distort the analysis or model training.

Data Transformation:

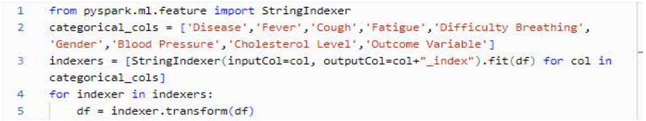
- Feature scaling: To stop some features from predominating over others, standardize or normalize numerical features to bring them to a similar scale.
- Coding categorical variables: Use methods like label encoding or one-hot encoding to translate category variables into a numerical format.
- Feature engineering: To improve the model's capacity to recognize patterns, add new features or alter current ones.

For removing null values the following state,ment is used in the dataset.



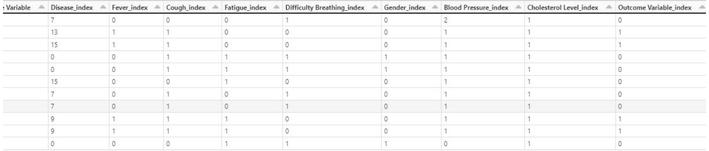
```
#Removing Null Values from the Dataset
df = df.dropna()
```

The data mainly contain Non Numerical values.To convert that types of values StringIndexer is used.For an example :



```
from pyspark.ml.feature import StringIndexer
categorical_cols = ['Disease','Fever','Cough','Fatigue','Difficulty Breathing',
'Gender','Blood Pressure','Cholesterol Level','Outcome Variable']
indexers = [StringIndexer(inputCol=col, outputCol=col+"_index").fit(df) for col in
categorical_cols]
for indexer in indexers:
    df = indexer.transform(df)
```

Here the disease in categorical value is converted to numerical value and the data looks as follows:



Variable	Disease_index	Fever_index	Cough_index	Fatigue_index	Difficulty Breathing_index	Gender_index	Blood Pressure_index	Cholesterol Level_index	Outcome Variable_index
	7	0	0	0	1	0	2	1	0
	13	1	1	0	0	0	1	1	1
	15	1	1	0	0	0	1	1	1
	0	0	1	1	1	1	1	1	0
	0	0	1	1	1	1	1	1	0
	15	0	0	1	0	0	1	1	0
	7	0	1	0	1	0	1	1	0
	7	0	1	0	1	0	1	1	0
	9	1	1	1	0	0	1	1	1
	9	1	1	1	0	0	1	1	1
	0	0	0	1	1	1	0	1	0

The converted values are added to the dataset as a new colums.

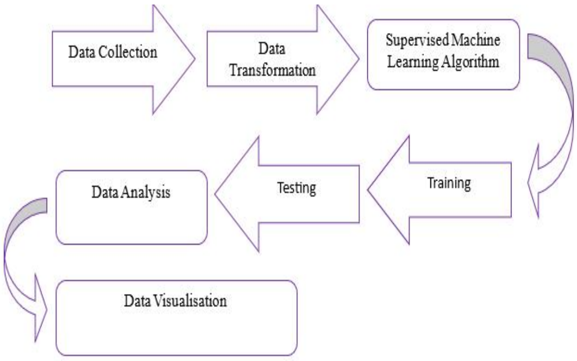


Figure 3 : Pipeline For Data Analysis.

.An exploratory data analysis (EDA) was conducted to understand the dataset's structure and relationships between variables. This involved displaying the initial rows and creating visualization of the general datas.

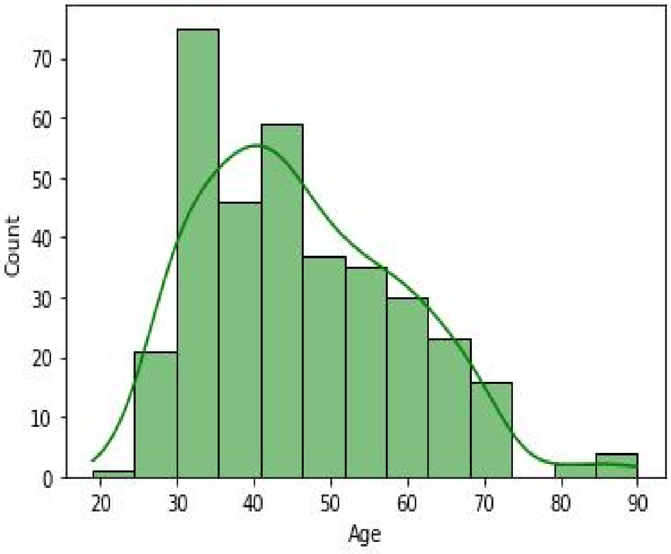


Figure 3.1 : General age distribution.

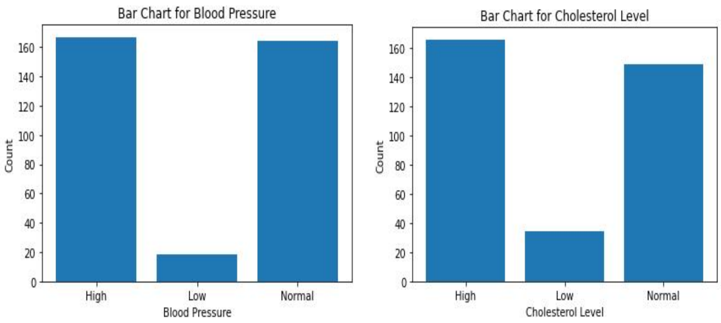


Figure 3.2 : Bar chart -Blood pressure and Cholesterol levels by age.

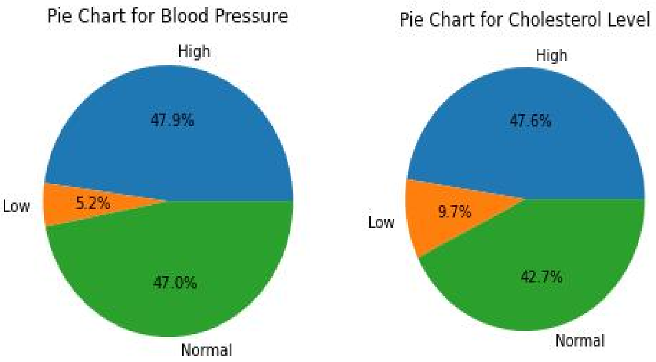


Figure 3.3 : Pie chart -Blood pressure and Cholesterol levels by age.

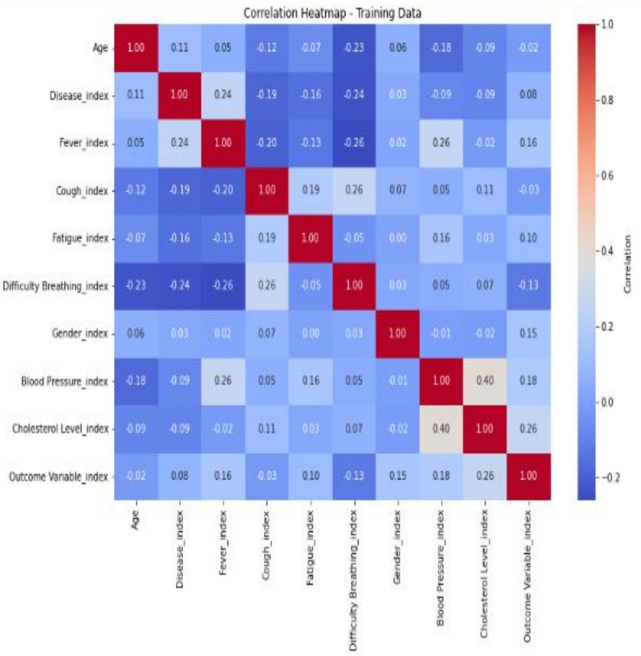


Figure 3.4 : Correlation heatmap-Training Data.

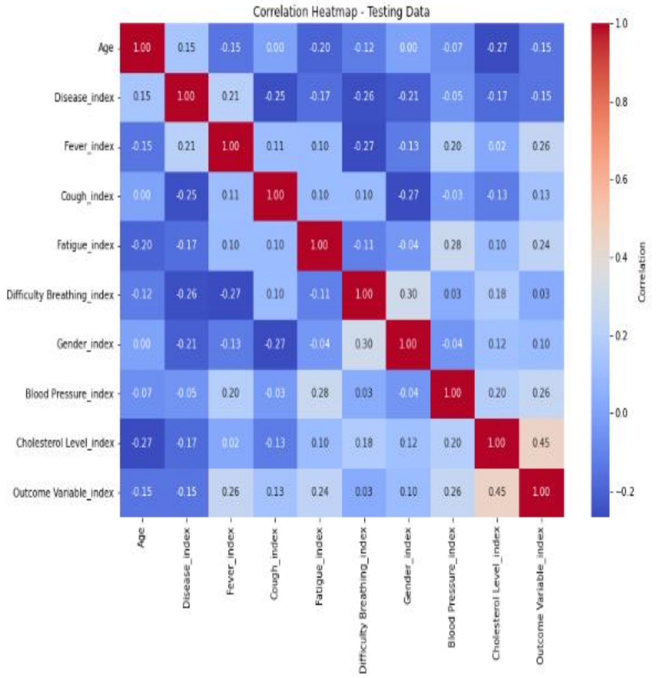


Figure 3.5 : Correlation heatmap-Testing Data.

C.Machine Learning Models

The following forecasting models were built and demonstrated..(a)LightGBM Classifier and (b)Random forest Classifier.

(a)LightGBMClassifier : An effective gradient boosting framework for distributed training on big datasets is called LightGBM (Light Gradient Boosting Machine). It was created by Microsoft and is frequently used for a range of machine learning applications, such as ranking, regression, and classification. LightGBM has a reputation for being extremely fast, efficient, and capable of handling big datasets.

(b)Random forest Classifier:The Random Forest classifier is an ensemble learning method that creates a lot of decision trees during training and produces a class that is either the mode of the classes (classification) or the mean prediction (regression) of the individual trees. This method is commonly used in machine learning when predicting a continues variable with complex relationships to avoid overfitting.

IV. RESULT

Based on the dataset provided, the result shows that the Disease Symptoms and Patient Profile Dataset can classify with both the classifiers.The LightGBMclassifier and Randomforestclassifier produces accurate and correct results.

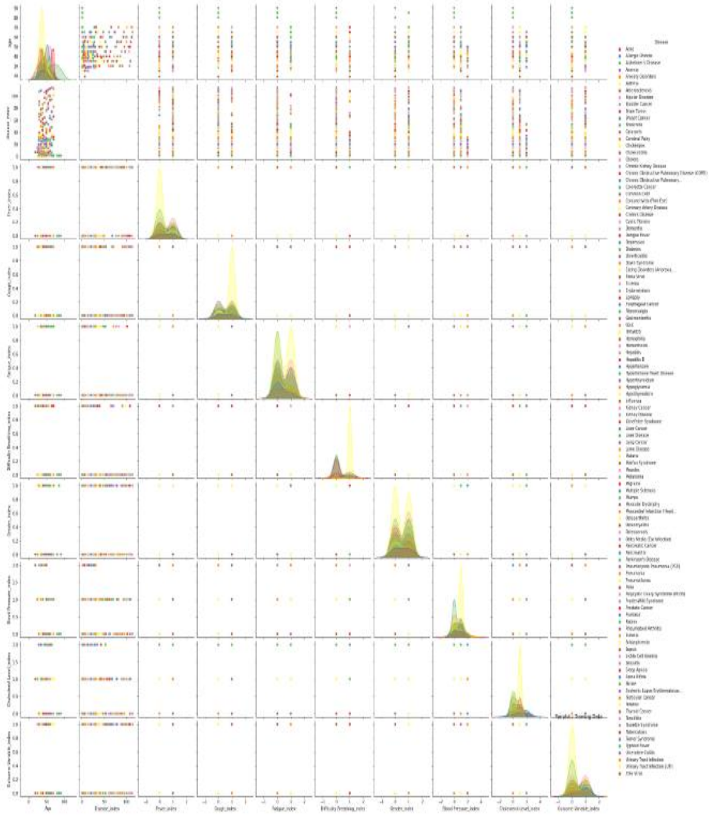


Figure 4 :Pairplot for the dataset.

We use 2 Predictive models for Comparison in this project evaluation to find which model is better.Fist we used the LightGBMclassifier model to classify the data and it have an accuracy of 73% .Then we use Randomforest classification ALgorithm.The Accuracy of Randomforest Classification is 80%.From the accuracy itself we can say that the Randomforest classification model is more accurate for Disease Symptoms and Patient Profile Dataset .

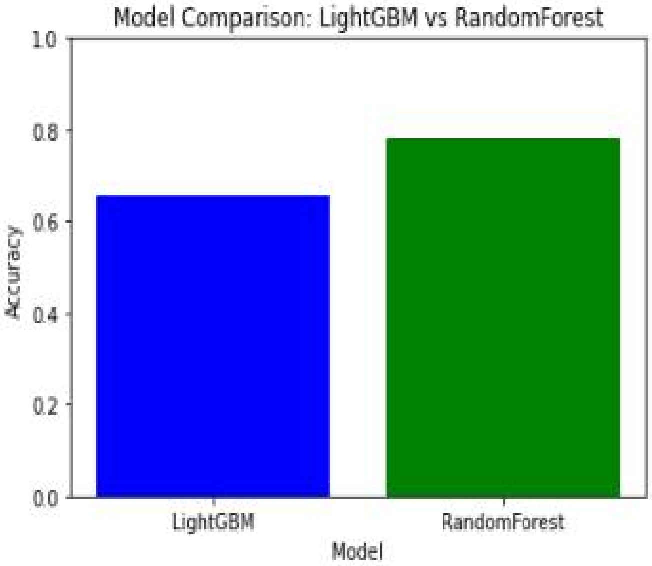


Figure 4.1 :Barchart for the Model Comparison.

From the Bar chart it is clear that the accuracy of Randomforest is more accurate than the LightGBM classifier.

The first goal is "For which disease the outcome variable is more positive?"

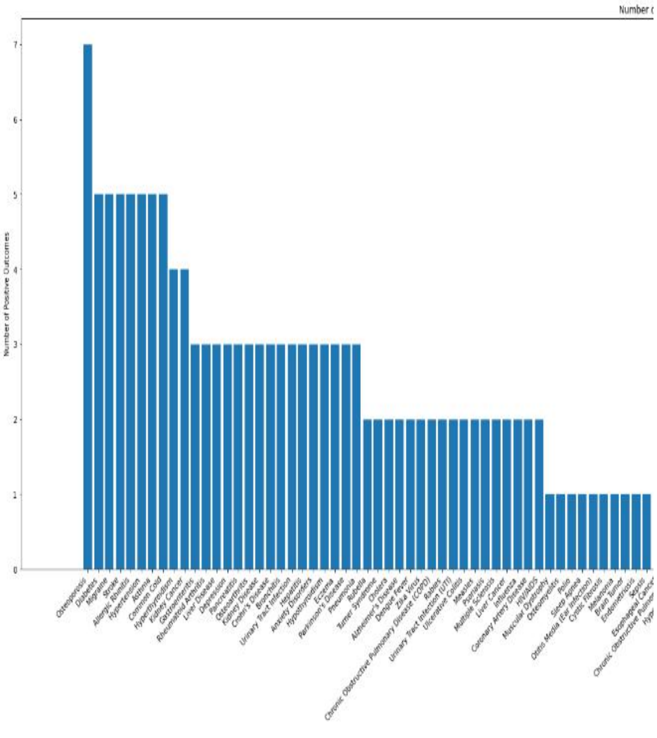
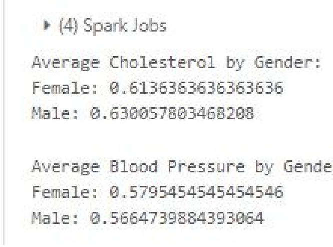


Figure 4.2 :Bar chart for the Positive Outcome Variables for Diseases

From the barchart it is clear that the first goal is success that the disease **"Osteoporosis"** has the highest positive Outcome Variable.

The Second goal is "For Which gender have high Cholesterol Level and high Blood pressure level.?"

For this we need to find the average percentage for high Cholesterol Level and high Blood pressure level for each genders(Female/Male).



```
▸ (4) Spark Jobs
Average Cholesterol by Gender:
Female: 0.6136363636363636
Male: 0.630057803468208

Average Blood Pressure by Gender:
Female: 0.5795454545454546
Male: 0.5664739884393064
```

Figure 4.3:Average Percentage Calculation.

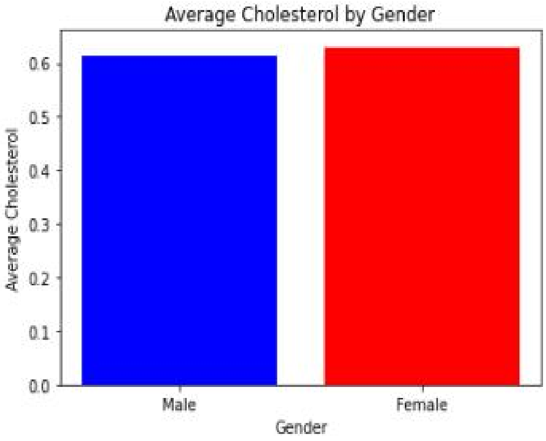


Figure 4.4:Average Cholesterol by Gender

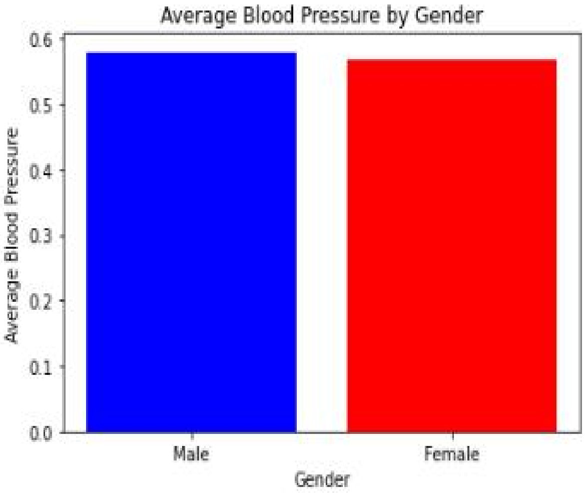
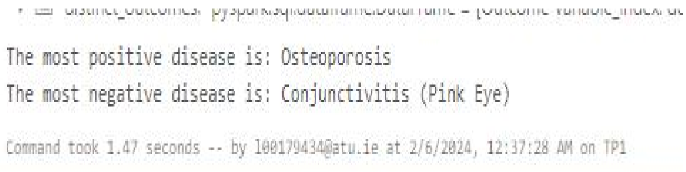


Figure 4.5:Average Blood Pressure by Gender

The Third goal is "Which disease has the most positive and negative Outcome Variable?"

For this we need to find the sum of positive and negative outcomes for each diseases and check distinct values of the outcome variables.Then create new variables to save the most positive disease and most negative disease and print those variables.



```
▸ ▤ distinct_outcomes: pyspark.sql.dataframe.DataFrame = [Outcome Variable_index: do

The most positive disease is: Osteoporosis
The most negative disease is: Conjunctivitis (Pink Eye)

Command took 1.47 seconds -- by 100179434@atu.ie at 2/6/2024, 12:37:28 AM on TP1
```

Figure 4.6 :Most positive and negative outcome diseases.

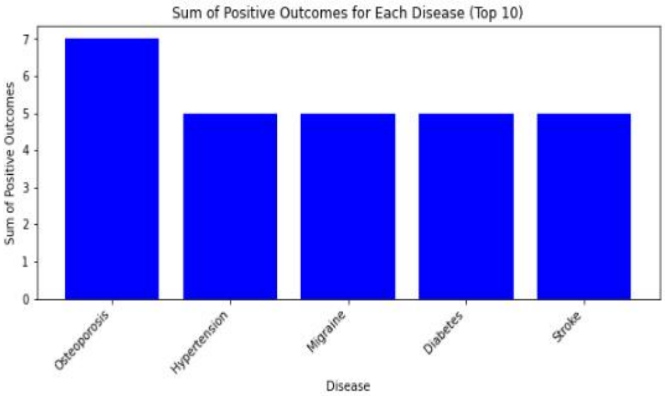


Figure 4.7:Bar chart of sum of positive outcomes.

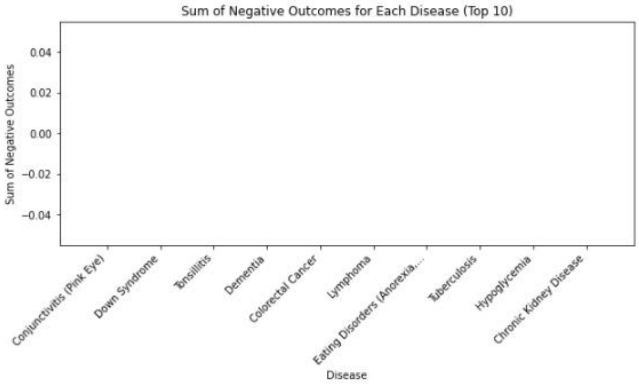


Figure 4.8:Bar chart of sum of negative outcomes.

V. CONCLUSION

This study proposes a method for analysis of the Disease Symptoms and Patient Profile Dataset with LightGBMclassifier and Randomforestclassifier.This is accomplished by employing the most powerful features available in the dataset.The user devised a solution by first processing the data and then extracting some dataset features using VectorAssembler in pyspark. The user then used the Classification models to create a model to analysis the dataset.

The work has a bright future ahead of it. The dataset can be enlarged and any other online learning algorithms or technologies can be used.Real-time identification of diseases in videos may be possible. Using graph theory and machine learning techniques, the other area of application is detecting important and different other symptoms of causes a diseases.

The following are the results of the research conducted during this study:

- The top features to analyze the dataset are "Disease , Fever , Cough , Fatigue , Difficulty Breathing, Gender , Blood Pressure,Cholesterol Level and the Labelled Data is Outcome Variable.
- The process that followed resulted in a 100% Accurate rate.
- The study of the categorical data encoding is very difficult, but in the case of pyspark, it would be Easier we can use StringIndexer for the data

conversion.

- The Randomforestclassifier Algorithm seemed to be the best algorithm for the analysis of the dataset, as it provided a higher accuracy rate and allowed for the assignment of a degree of confidence to each piece of information.
- The first goal is “For which disease the outcome variable is more positive?”.The Outcome Variable in the Dataset **“Disease Symptoms and Patient Profile ”** is more positive is for the Disease **“Osteoporosis”.**
- The Second goal is “For Which gender have high Cholesterol Level and high Blood pressure level.?”.The Cholesterol level is higher for the gender **“Male”** and The Blood Pressure is higher for the gender **“Female”.**
- The Outcome Variable in the Dataset **“Disease Symptoms and Patient Profile ”** is more positive for the Disease **“Osteoporosis”** and is more negative for the Disease **“Conjunctivitis (Pink Eye)”**

VI. REFERENCES

[1] Kaggle ,Datasets ,Disease Symptoms and Patient ProfileDataset:https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset.

[2] scikit-learn Machine Learning in Python,Tutorials : https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[3]Apache spark , Spark Machine Learning Library (MLlib)Guide :https://spark.apache.org/docs/1.4.1/mllib-ensembles.html#random-forests.

[4]UnitedweTech,bulitin,Randomforest:https://builtin.com/data-science/random-forest-python